AFRL-RH-WP-TR-2009-0033

Intelligibility of Target Signals in Sequential and Simultaneous Segregation Tasks

Nandini Iyer
Douglas S. Brungart
Brian D. Simpson
Warfighter Interface Division
Battlespace Acoustics Branch

March 2009

Final Report for October 2004 to March 2008

Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Warfighter Interface Division
Battlespace Acoustics Branch
Wright-Patterson AFB OH 45433

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2009-0033 HAS BEEN REVIEWD AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR

//signed//
Douglas S. Brungart, Ph.D.
Program Manager
Battlespace Acoustics Branch

//signed//
Daniel G. Goddard
Chief, Warfighter Interface Division
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
25-03-2009	Final	1 October 2004 – 1 March 2008

4. TITLE AND SUBTITLE
Intelligibility of Target Signals in Sequential and Simultaneous Segregation Tasks

5a. CONTRACT NUMBER
In-House

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Nandini Iyer, Douglas S. Brungart, Brian D. Simpson

5d. PROJECT NUMBER
2313

5e. TASK NUMBER
HC

5f. WORK UNIT NUMBER
2313HC52

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
711TH Human Performance Wing
Human Effectiveness Directorate
Warfighter Interface Division
Battlespace Acoustics Branch
Wright-Patterson AFB OH 45433-7901

10. SPONSOR/MONITOR'S ACRONYM(S)
711 HPW/RHCB

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RH-WP-TR-2009-0033

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
88 ABW/PA Cleared 04/07/09; 88ABW-09-1373.

14. ABSTRACT
Two experiments are described in the report: the first experiment investigated target intelligibility in a sequential segregation task, while the second experiment investigated performance in a simultaneous segregation task. In the first experiment, speech intelligibility was examined in the presence of four types of maskers (continuous noise, interrupted noise, continuous speech and interrupted speech) at seven interruption rates. Results highlight the important role that target-masker similarity plays in the segregation of sequentially-presented speech signals. In the second experiment, we assessed whether listeners could extract information from target phrases that were masked either by contextually similar maskers that were confusable with the target or contextually irrelevant maskers where no such confusion was possible. The results show that a large multimasker penalty occurs when at least one of the interfering talkers is contextually similar to the target, but that no multimasker penalty occurs when neither of the masking voices is confusable with the target. These results suggest that the primary impact of the multimasker penalty is a severe degradation in the listener's ability to link together words that occur at different points in the target phrase.

15. SUBJECT TERMS
Informational masking; energetic masking, multimasker penalty, speech perception

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Nandini Iyer
UNCL	UNCL	UNCL	SAR	41	19b. TELEPONE NUMBER (*Include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI-Std Z39-18

THIS PAGE INTENTIONALLY LEFT BLANK.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

In everyday listening environments, a listener must often separate a target speech stream from one or more streams of irrelevant or interfering speech. In such situations, the target and the irrelevant streams might overlap in time, which gives rise to the difficult task of segregating simultaneous speech sounds. In many cases, there is only minimal overlap of target and irrelevant signals in time. In such instances, it can still be difficult to follow the target stream, and target intelligibility will depend on the listeners' ability to group sequential elements of the target together. In the first experiment, speech intelligibility was examined in the presence of four types of maskers (continuous noise, interrupted noise, continuous speech and interrupted speech) at seven interruption rates. With a noise masker, listeners performed significantly better when the target and masking signals were alternated than when both were on continuously. However, with a speech masker, performance was worse in the condition where the masker was interrupted than when it was continuous. Further, performance with a speech masker improved, rather than decreased, when the interrupted target was paired with a continuous masker. Introducing a pitch difference between the target and masker improved intelligibility in the interrupted masker condition, and exceeded performance in the continuous masker condition at some interruption rates. The results further highlights the important role that target-masker similarity plays in the segregation of sequentially-presented speech signals.

In the second chapter, two experiments were conducted to assess whether listeners could extract information from target phrases that were masked either by contextually similar maskers that were confusable with the target or contextually irrelevant maskers where no such confusion was possible. The results show that a large multimasker penalty occurs when at least one of the interfering talkers is contextually similar to the target, but that no multimasker penalty occurs when neither of the masking voices is confusable with the target. Remarkably, in cases where one interfering talker is contextually similar to the target, the similarity of the second interfering talker has little or no impact on performance. Indeed, when the results were corrected for random guessing, there was no difference in performance between conditions with three contextually confusable talkers and those with two confusable talkers and one irrelevant talker. These results suggest that the primary impact of the multimasker penalty is a severe degradation in the listener's ability to link together words that occur at different points in the target phrase.

THIS PAGE INTENTIONALLY LEFT BLANK

1.0 Introduction

Two mechanisms are generally implicated in sound source segregation: simultaneous and sequential sound source segregation. The former refers to listeners' ability to "hear out" a target signal presented in the midst of competing interfering signals. The latter refers to processes that allow listeners to group together target segments when that are interleaved with segments of the interfering sound. And while it seems obvious that the ability to simultaneously segregate a target from masking sounds is important in order to understand speech in noise, the role of sequential segregation is less clear. Recently, Mackersie et. al., (2001) demonstrated a strong correlation between speech reception thresholds (SRT) and tonal fusion thresholds (Rose and Moore, 1997), suggesting that sequential segregation might also be important in sound source segregation.

This report describes two experiments that investigated both simultaneous as well as sequential sound source segretation. The report is divided into two chapters: each chapter presents the relevant findings from two experiments. The first experiment describes the differential effects on target intelligibility when a target speech signals is interleaved with either a speech or a noise masker. The second experiment describes a simultaneous segregation task and addresses the masking efficiencies of different types of speech and noise maskers. It also introduces a method of assessing the importance of the call-sign in the CRM-based intelligibility test, and way to correct for it.

2.0 Sequential segregation: Effects of periodic masker interruption on the intelligibility of speech

Auditory scenes are easiest to process when they contain only a single source, but in the real world sounds rarely occur in isolation. Consequently, listeners are often required to integrate meaningful audio signals from the loose collections of spectro-temporal "glimpses" that emerge when a target signal has a different pattern of spectro-temporal energy fluctuations than the masker. Less frequently, listeners might be required to process audio signals that are "interrupted" by brief periods of silence (e.g., from a poor communications line such as a cell-phone). In either case, the listener is faced with the somewhat daunting challenge of identifying the spectro-temporal fragments that should be grouped with the target signal, segregating those fragments from any interfering sounds that might be present in the auditory mixture, and using those fragments to somehow reconstruct the original sound source.

While in many ways this might seem like a difficult or impossible task, a number of laboratory experiments have demonstrated that listeners can do it with relative ease, particularly when the target signal happens to be speech (Miller and Heise, 1950; Howard-Jones and Rosen, 1993; Buss et al., 2004). Miller and Licklider (1950) were among the first to demonstrate that intelligibility of a target sentence is only minimally affected when it is periodically interrupted by silence. When the rate of interruption is 4 Hz or less (i.e., the periods of silence are relatively long), there is some loss of information because entire syllables and words are eliminated from the stimulus. However, once the rate of interruption reaches the point where there are multiple glimpses of the target speech per phoneme (between 8 - 100 Hz), the interrupted words become just as intelligible as uninterrupted speech.

Listeners are also quite adept at integrating target information across temporal intervals that are either alternated with or masked by noise. For example, at 0 dB signal-to-noise ratio (SNR), word intelligibility is not significantly altered when continuous speech is intermittently masked by white noise (Miller and Heise, 1950; Dirks and Bower, 1969, 1970) with interruption rates varying from 1 Hz to 100 Hz; in fact, at these rates of masker interruption, word intelligibility is only slightly worse than it is for speech that is interrupted by silent intervals. These results demonstrate that high levels of intelligibility can be achieved even when the listeners only have access to brief glimpses of a target speech stimulus. Indeed, this phenomenon has been demonstrated in numerous masking experiments, where target intelligibility has been significantly improved by interrupting or amplitude modulating a noise masker (Miller, 1947; Pollack, 1955; Wilson and Carhart, 1969; Dirks and Bower, 1970). In fact, listeners are so good at attending to speech that is interrupted by noise that they often report hearing it as a continuous speech signal masked by interrupted noise, rather than as an interrupted speech signal alternating with interrupted noise (although intelligibility is not improved relative to the condition where the speech is interrupted by silence) (Miller and Heise, 1950). The fact that listeners perceive an auditory signal "continuing" through a more intense burst of masking noise is related to a well-researched perceptual phenomenon known variously as *picket fence* effect, the *temporal induction* effect, the *phonemic restoration* effect, or the *continuity illusion*, and its effects have been documented in a number of

studies (Thurlow and Elfner, 1959; Dirks and Bower, 1970; Warren, 1970; Miller and Heise, 1950; Warren et al., 1972; Drake and McAdams, 1999).

Although previous experiments have thoroughly explored situations where an interrupted speech signal is masked by, or alternated with a noise masker, virtually no information is available regarding what might occur when an interrupted speech signal is masked by or alternated with a *speech* masker. Previous experiments have shown that speech maskers can impact speech intelligibility much differently than noise maskers in a wide variety of different listening configurations. For example, listeners appear to be able to improve performance by listening for the quieter of two talkers in a multitalker mixture (Brungart, 2001; Brungart et al., 2001), but they gain no benefit from listening for a quiet talker in noise. Similarly, listeners have been shown to benefit more from real or apparent spatial separation between a speech target and a speech masker than from real or apparent spatial separation from a noise masker (Hawley et al., 2000; Arbogast et al., 2002; Freyman et al., 1999). These differences, which are fundamentally related to the listener's inability to distinguish the acoustic features of a target speech signal from those of a similar-sounding speech masker, have often been referred to as "informational masking" (Freyman et al., 1999; Kidd et al., 1998; Brungart, 2001; Brungart et al., 2001). Whereas energetic masking refers to the masking that occurs to the overlapping target and masker energies in the peripheral auditory filters, the term informational masking has been commonly used to describe the difficulty that listeners experience when attempting to separate elements of a target signal from that of a perceptually similar masker.

In light of the many prior experiments that have found different masking patterns for speech and noise maskers, one might anticipate that different results would be obtained when a target speech signal is alternated with a competing speech signal than when it is alternated with a competing noise signal. In both instances, the listener is afforded glimpses of the target signal where SNR is close to optimal, and from the standpoint of energetic masking this would lead to the prediction that there might be a significant improvement in performance compared to the condition when the target and masker are both continuous signals. This result would be expected with a noise masker, because it is perceptually quite easy to segregate an "interrupted speech signal" from an "interrupted noise signal." Thus, one might argue that nearly all the masking that occurs with noise maskers is energetic masking that will be reduced when the target and masker are presented in alternating intervals rather than simultaneously. However, the same cannot be said for a speech masker. From an informational masking point of view, alternating a speech masker with a target signal would leave the target and masker *similar* in the sense that both fall in the category of "interrupted speech signals." Thus, we would expect little or no release from informational masking in this case. On the other hand, if the target is interrupted and the masker is continuous, then listeners could potentially utilize a "continuity" cue to identify the interrupted target and/or ignore the continuous masker, thus decreasing the effects of informational masking and improving performance. It is also worth noting that in cases where target and masking speech signals are more easily distinguishable based on characteristics other than than their pattern of temporal interruption (such as differences in talker F0), the effects of informational masking would generally be reduced. In some cases, the alternation might lead to some improvement

in performance due to the reduced energetic masking in that condition.

The aim of this experiment was to determine if type of masker (speech or noise) and its continuity (interrupted or continuous) does indeed influence the identification of target speech signals. In addition, the experiment examined the extent to which differences in voice characteristics of the target and masker in the speech masking conditions contributed to target identification.

2.1 Experiment 1: Effect of types of masker and interruption rates

2.1.1 Methods

2.1.1.1 Listeners

Eight listeners, ranging in age between 21 and 55 years, with normal hearing thresholds (audiometric thresholds $\leq$15 dB HL from 250 Hz to 8000 Hz) participated in the study. All were paid volunteers and had previous experience participating in experiments utilizing the speech corpus employed in the current study.

2.1.1.2 Speech Materials

The speech materials were sentences from the publicly available Coordinate Response Measure (CRM) corpus (Bolia et al., 2000). The corpus contains recordings of four male and four female talkers speaking phrases of the form, "Ready [call-sign] go to [color] [number] now." A combination of eight call-signs (Arrow, Baron, Charlie, Eagle, Hopper, Laker, Ringo, and Tiger), four colors (red, white, blue, green) and eight numbers (1-8) resulted in 256 phrases recorded per talker for a total of 2048 phrases in the entire corpus. The sentences in the corpus were band-limited to 8 kHz and sampled at 40 kHz. In this experiment, the target sentence was always denoted by the call-sign "Baron." Listeners heard the target sentence in the presence of four types of maskers: continuous noise, interrupted noise, continuous speech or interrupted speech. The noise masker was a Gaussian noise that was spectrally shaped so that it matched the long term average spectrum of the 2048 phrases in the CRM corpus and was presented at a RMS level that was 8 dB more intense than the RMS level of the target speech (i.e., at an SNR = -8 dB) prior to any interruption. When the masker was speech, a CRM phrase with a call-sign other than "Baron" was used and a different color and number than the target phrase was selected randomly from the corpus, set to have the same RMS level as the target speech, and played simultaneously with the target phrase. In order to maximize similarity between the voice characteristics of the target and masking sentences, the masking talker was always the same talker as the target talker.

The target sentence was interrupted in all the experimental conditions. This interruption was achieved by multiplying the target sentence with a square wave with a 50 percent duty cycle and repetition rates of 4, 8, 16, 32, 64, or 128 Hz. The target sentence was always multiplied by the square wave with an "on" starting phase. When the masker was continuous, its onset was simultaneous with the interrupted target sentence. In conditions where the masker was also interrupted, it was multiplied by a square wave and switched at the same rate as the target, but with the opposite starting phase (i.e., the "off" phase). The

interrupted target and masking signals were then summed together and presented to the listeners. Thus, in experimental conditions where the masker was interrupted, the resulting signal consisted of alternating segments of target followed by masker segments. A control condition, where the target and masker were both continuous, was also presented to listeners, and for convenience we identified that condition as the "0 Hz" condition. To summarize, there were two independent variables in the experiment: interruption rate (0, 4, 8, 16, 32, 64 or 128 Hz) and masker type (continuous noise, interrupted noise, continuous speech, or interrupted speech).

Figure 1 illustrates the experimental condition in which the target sentence (depicted in black) and a speech masker (depicted in gray) were interrupted at a rate of 4 Hz. The middle panel depicts the square wave with which the target and masker sentences were multiplied. As a result, listeners heard the target phrase interspersed by the masker phrase at a rate of 4 Hz. Note that the two square waves for the target and masker have identical repetition rates but different starting phases.

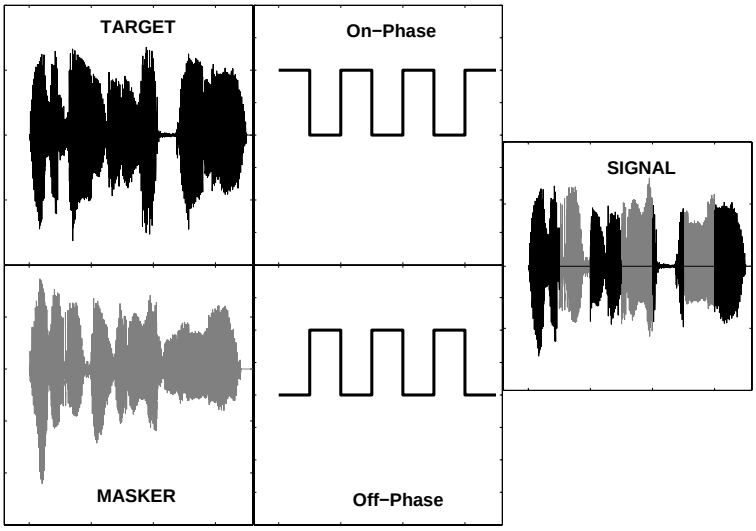

Figure 1: Representation of the signals presented to listeners in Experiment 1 in the condition where the target and speech masker alternated with each other at the rates of 4 Hz. Two phrases from the CRM corpus (target waveform depicted in black and masker waveform in gray) are multiplied by a square wave, such that the target was multiplied by the "on" starting phase (top middle panel) and the masker, by the "off" phase (bottom middle panel). The resulting signal is segments of target phrase that alternates with those of the masker phrase.

2.1.1.3 Procedure

The sentences from the corpus were presented diotically through Beyer DT 990 PRO headphones to listeners seated in a quiet listening room. Potential responses were displayed in a 4x8 matrix of colored digits. Listeners responded by using a cursor to select the desired

color-number combination in the target sentence. Correct response feedback was provided after every trial, and overall performance for each subject as well as group averages were displayed at the end of each block. In a block of trials, the masker type remained fixed, and four repetitions of each interruption rate were presented in a random order. The order of presentation of each masker type was also randomized across subjects. An additional block of trials was also conducted where listeners heard the target interrupted at rates of 4, 8, 16, 32, 64, and 128 Hz with no masker. A total of 20 trials was collected per masker type at each interruption rate for each subject.

2.1.2 Results and Discussion

Figure 2 shows the percentage of correct color and number responses as a function of interruption rate for the continuous and interrupted noise masker conditions (left panel) and the continuous and interrupted speech masker conditions (right panel). The data were averaged across all eight listeners and the error bars represent the 95 percent confidence intervals. The unfilled triangles in both panels represents performance in the control conditions, where listeners heard a continuous target in the presence of a continuous masker (the 0 Hz condition). The circles represent the performance in the continuous masker condition, whereas the squares represents performance in the alternating masker condition. The diamonds represent the data from the condition where listeners heard the target alone interrupted at rates of 4, 8, 16, 32, 64, and 128 Hz. The left and right panels depict the data for noise and speech maskers, respectively.

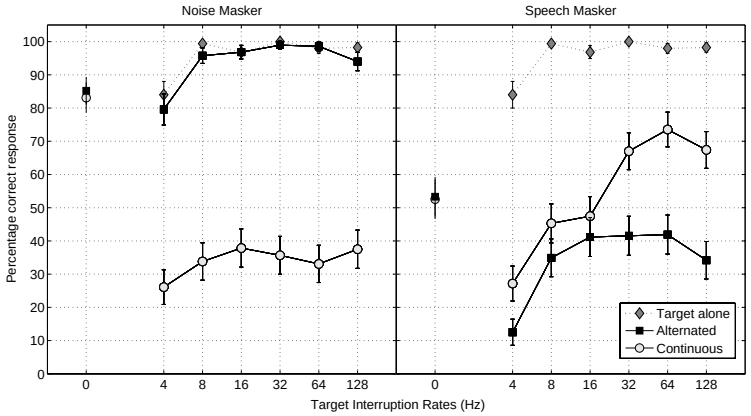

Figure 2: Percent correct target sentence identification as a function of the interruption rate for four masker types. The circles represent performance for listeners in conditions where the masker was continuous and the squares represent performance when the maskers were alternated with the target. The diamonds represents performance in the interrupted target condition with no masker. The left panel plots the results for noise maskers whereas the right panel plots those for speech maskers. In both panels, the unfilled triangles represents data for 0 Hz control condition

In the noise control condition, where the target and masker were both on continuously (0 Hz point in the left panel of the figure), the listeners responded with the correct color

and number combination in approximately 85 percent of the trials. When the target was interrupted but presented with no masker (diamonds), performance was near 100 percent correct at rates of 8 Hz and above. These results are in good agreement with those from Miller and Licklider (1950) using monosyllabic words from the phonetically balanced (PB) list.

Performance with the noise masker remained near ceiling when the target was alternated with a noise masker, suggesting that the introduction of an alternating noise had no substantial impact on intelligibility with an interrupted speech masker. These results differ somewhat from those reported by Miller and Licklider (1950), where an alternating noise had a significant influence on intelligibility when the interruption rates exceeded about 10 Hz. These differences could be attributed to the fact that the CRM is generally more robust to interference due to noise than the PB 50 list. For example, Brungart (2001) reported that the intelligibility scores obtained using the CRM showed asymptotic performance with a speech-shaped noise masker when SNR exceeded -6 dB or an AI value of 0.25. On the other hand, PB word lists are substantially more sensitive than the CRM at AI values greater than .25 (Kryter, 1969). In other words, the CRM words had more phonetic redundancies than the PB words, and thus their identification was more tolerant to the introduction of interrupted noise.

In contrast to the results with an alternating noise masker, target intelligibility for an interrupted speech signal in the presence of a continuous noise masker (circles in the left panel of the figure) were significantly reduced relative to the continuous control condition at all interruption rates tested. Summarizing the results from the listening conditions where the masker was noise, data showed that the listeners were able to obtain enough information to identify 85 percent of the color-number combinations when they heard the entire target signal at an SNR of -8 dB; 100 percent of the color-number combinations when they heard 50 percent of the target signal at an SNR of ∞ dB; and roughly 35 percent of the color-number combinations when they heard 50 percent of the target signal at an SNR of -8 dB.

Performance with speech maskers showed a remarkably different trend than noise maskers. First, note that performance was significantly lower in the 0 Hz speech control condition (continuous target, continuous speech masker) than in the 0 Hz noise condition despite the nominally higher signal to noise ratio (0 dB versus -8 dB). This difference is consistent with previous results obtained in two talker conditions (Brungart, 2001), and may be attributable to informational masking (i.e., the inability of listeners to separate the two CRM phrases into distinct utterances).

Third, note that, also in direct contrast to the noise condition, performance in the continuous masker condition (circles in Figure

In order to verify these results, the arcsine-transformed percentage correct scores were submitted to a two-factor within subject ANOVA with type of masker and interruption rates as factors. The analysis indicated that there was a significant main effect of interruption rate ($F_{(6,42)}=59.21$, $p<0.001$) and masker type ($F_{(3,21)}=136.78$, $p<0.001$). There was also a significant interaction between interruption rate and masker type ($F_{(18,126)}=15.27$, $p<0.001$) and it reflects the fact that at interruption rates of 32 Hz and higher, there is a release from masking with a continuous speech masker but not for the continuous noise masker.

The results from Experiment 1 indicate that the percentage of correct color-number identifications for a target sentence in the CRM corpus is largely dependent on the type and continuity of the masker. For noise maskers, the results obtained concur with those reported in prior studies; in particular, target word identification improves when the target alternates with the masker but deteriorates with interrupted target and continuous masker. The finding that an alternating noise masker is less effective than a continuous noise masker is well-documented in the literature (Miller and Heise, 1950; Dirks and Bower, 1969; Howard-Jones and Rosen, 1993). The release from masking is attributed to instances in time when the local signal-to-noise ratio is greatly improved. The auditory system is fairly adept at distinguishing the noise fragments from the speech fragments and integrating segments of the target sentence together so that performance is no different from the condition when the target sentence is interrupted by silence. In conditions when an interrupted target is heard in the presence of a continuous masker, three factors may be responsible for the degradation in performance from the control condition. First, the listener is not afforded a glimpse of the target sentence at a favorable signal-to-noise ratio that is sufficiently long to integrate target information. Second, the listener loses 50 percent of the information in the target speech due to the interruption. And third, the listener faces the potentially much tougher problem of extracting "noisy" speech fragments from noise (rather than extracting "clean" speech fragments from noise) in order to piece together the fragments of the interrupted target signal into a coherent stream. From the data it is evident that listeners failed to "fuse" the temporally separated segments of the target at all rates of interruption tested in this experiment.

Although they cannot be directly compared to any previous experiments, the results obtained with the speech masker were in fact much different from those obtained with the noise masker. In general, these results were consistent with the predictions that were laid out in the introduction. The listeners' performance in the alternating speech masker conditions was consistently worse than with continuous speech masker. From the standpoint of energetic masking alone, this result is surprising because in the alternating speech masker condition the local signal-to-noise ratio is just as favorable as in the alternating noise masker condition. However, there is ample evidence that informational masking effects related to the confusability of the target and masking voices play a much greater role than energetic masking in determining performance in the two-talker CRM task when the target and masking voices are spoken by the same talker. Also, apart from the use of a strategy where a listener is able to synchronously listen for a target only in those intervals where the target speech is present in the stimulus, there is no reason to believe that alternating the target and masking voices will produce any significant decrease in informational masking. However, significant increases in target identification performance were observed when target speech was masked by a continuous rather than an alternating speech masker, suggesting that decreasing the perceptual similarity of the target and masker by making the masker "continuous" can result in a reduction in informational masking. While we are unaware of any previous results demonstrating this effect with a speech signal, it is worth noting that a number of studies have reported similar effects with tonal non-speech signals that also tend to generate significant amounts of informational masking. For example, Kidd et. al., (1994) reported

substantial reduction in informational masking when a 1000 Hz target was presented during every alternate instead of in every burst of a multitone masker. This reduction in masking was obtained despite the fact that the target energy in the alternating condition was half that present during the continuous condition. Similarly, Neff (1995) reported that informational masking was substantially reduced for an amplitude-modulated (AM) target compared to a baseline pure-tone target presented with multi-component maskers.

One important question left unanswered by these data is whether the listeners in the speech masking condition actually received *any* benefit from regions of favorable SNR that occurred in the alternating speech condition. If energetic masking was not a factor in the speech masking conditions, then one might expect to get the same results regardless of whether the interrupted target and masking signals were presented in the same time intervals (i.e. both on, both off) or in alternating time intervals (i.e. target on, masker on). In order to address this question, a follow-up experiment was conducted that compared the alternating condition tested in Experiment 1 with a new condition where the target and masking signals were both interrupted, but were presented simultaneously. The experimental procedures were identical to those used in Experiment 1, but data were only collected at four interruption rates: 0 (control condition), 8, 32 and 128 Hz. Figure 3 presents the average results from seven listeners, six of whom were also in Experiment 1. As in Figure 2, the circles and squares represent the average percentage correct color-number responses for a continuous and interrupted masker respectively. The diamonds represent average performance in the condition where the masker was pulsed on and off simultaneously with the target signal. Comparing the simultaneous speech masking condition (diamonds in the right panel of the figure) to the alternating speech masking condition (squares), it is clear that some release from energetic masking was obtained when the speech masker was interleaved with the target rather than presented simultaneously. However, the improvement in performance was small compared to that obtained with the continuous masker (circles in the figure), despite the fact that the continuous masker condition had the same local SNR value as the simultaneous masking condition. Perhaps a more surprising finding is that performance in the noise masking condition was also worse in the simultaneous masking condition (diamonds in the left panel of the figure) than in the continuous masking condition (circles), despite the possibility of additional forward or backward masking from the noise in the target-silent intervals. The reason for the improvement is not clear, but one interesting possibility is that listeners were able to use the continuity of the noise masker to help segregate the noise from the target, in the same way they were able to use the continuity cue with the speech masker.

2.2 Experiment 2: Effect of changing voice characteristics

The results from the first experiment clearly indicate that the masking obtained with a continuous or interrupted speech masker was drastically different from a continuous or interrupted noise masker. In particular, target identification was significantly worse with an interrupted masker than with a continuous speech masker, whereas the opposite was true with a noise masker. In part, we hypothesized that the results were due to the similarity between the target and the masker. In all listening conditions in Experiment 1, the target

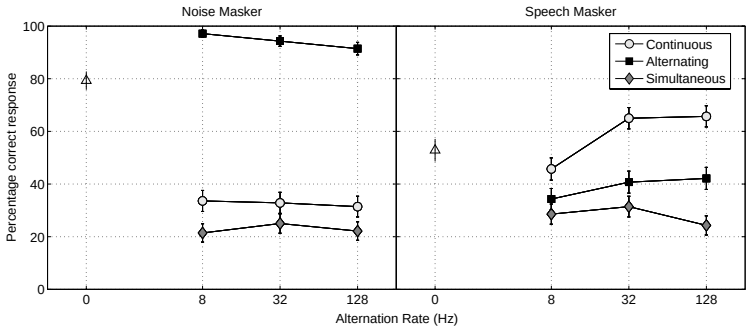

Figure 3: The percentage of correct responses plotted as a function of the interruption rate (Hz). The left panels depicts data from seven listeners with a noise masker, whereas the right panel depicts those with a speech masker. The circles and squares represent data for continuous and interrupted maskers respectively, whereas the diamonds represent data in the listening condition where the maskers were simultaneously presented with the interrupted target signal.

talker was always the same as the masking talker. In the second experiment, we varied the pitch differences between the target and masking talker for the speech masking conditions. If the similarity between the target and masker voice characteristics was a factor, then we would expect to see differences in target identification based on the amount of distinction between the voices.

2.2.1 Method

The target signals in this experiment were similar to those in Experiment 1. The masker was always a speech masker and was either continuous or interrupted. In addition to the masker continuity (i.e., continuous vs. interrupted masker), the masker sentences were processed using the Pratt software (Boersma and Weenink, 1996) implementation of the PSOLA algorithm (Moulines and Charpentier, 1990), so that the fundamental frequency (F0) and the vocal tract (v.t.) changes simulated a change from a masker that was the same sex as the target to one that was of the opposite sex. In particular, four manipulations of the F0 and v.t. produced masker sentences that varied from the target sentence by 2.7, 5.2 and 12.4 semitones. For a male target talker, the manipulation was akin to playing a masker phrase that can be described as 'quarter-female', 'half-female', and 'super-female' talker respectively. In order to retain the same semitone difference, absolute F0 and v.t. values varied for male or female voices. These values and additional details about the F0 and v.t. manipulations are described in Darwin et. al., (2003). Thus, there were three independent variables in this experiment: masker type (continuous vs. alternating speech masker), interruption rates (4, 8, 16, 32, 64, or 128 Hz) and pitch difference between the target and masker voice (2.7, 5.2 or 12.4 semitones). Data were collected from nine listeners with hearing thresholds within normal limits. Eight of the nine listeners also participated in the previous experiment. The interruption rate as well as pitch difference were randomized, but only one masker type was tested in each block. Data from the experiment represents

the average from 20 trials per condition per listener.

2.2.2 Results and Discussion

Figure 4 shows the percentage of correct color and number identifications as a function of interruption rates in each of the three pitch separation conditions. For comparison, data with the speech maskers from Experiment 1 are also plotted, where the masking talker was the same as the target talker and hence the overall pitch difference was 0 semitones. Within each panel, the gray circles show performance for the continuous speech masker, whereas the black squares show performance for a speech masker alternating with the target phrase. As in the previous experiment, the unfilled triangles represent data from the 0 Hz control condition.

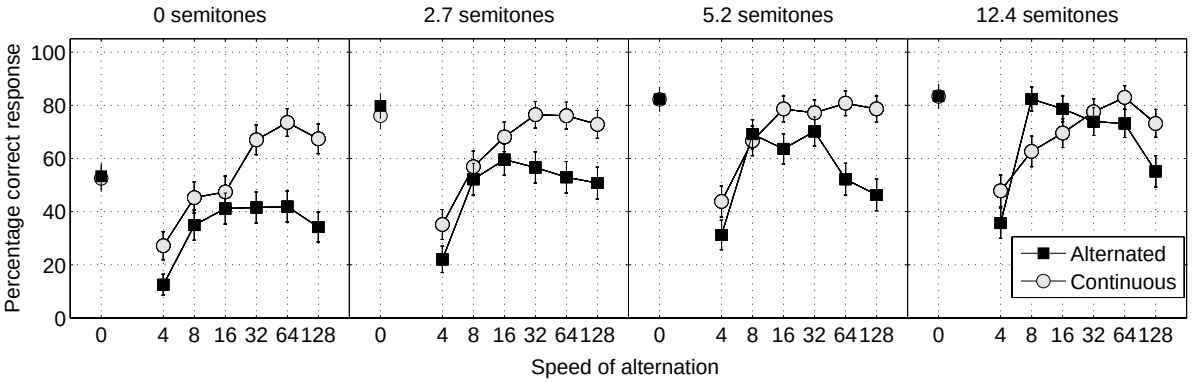

Figure 4: The percentage of correct responses plotted as a function of the interruption rate (Hz). The four panel represents performance of the listeners in experimental conditions where the masker and the target differ by 0, 2.7, 5.2 or 12.4 semitones. The results with continuous maskers are depicted in gray circles, whereas those with the alternating maskers are depicted in black squares. Data from 0 Hz control condition are plotted as unfilled triangles.

Target identification in the control condition (0 Hz) with a continuous target and a continuous masker (leftmost point in each panel) improved significantly when the pitch difference was increased from 0 to 2.7 semitones, but differences in pitch that exceed 2.7 semitones did not provide any additional benefit to listeners. These results are consistent with (Darwin et al., 2003), where target identification improved by 15-25 percent as the pitch difference increased from 0 to 2.7 semitones but plateaued at higher pitch differences.

In the interrupted-target conditions of the experiment, it is interesting to note that the introduction of a pitch difference had a much smaller effect in the continuous masker conditions of the experiment (circles in the figure) than in the alternated-masker conditions of the experiment (squares in the figure). In the continuous-masker conditions, the introduction of a pitch difference resulted in some increase in performance at low interruption rates (4-16 Hz), but had almost no effect at high interruption rates. Notably, performance in the higher interruption rates ($\geq$32 Hz) tended to plateau in all cases at the same level (80 percent

correct responses) achieved in the continuous 0 Hz control conditions when there was a pitch difference in the signal. A reasonable explanation for this result is that the pitch difference cue and the "continuous versus interrupted" cue were each equally effective at allowing the listener to distinguish the target speech from the masking speech, but that the cues were redundant rather than additive in cases where both were available to the listener. In other words, the listeners could distinguish the target and masking speech signals on the basis of pitch or on the basis of continuity, but they gained no further advantage when both cues were available.

The introduction of a pitch difference cue had a much larger impact on performance when the target and masker were alternated (filled squares in the figure). In these conditions, the introduction of a pitch difference had the greatest effect at intermediate interruptions rates (8-32 Hz), where it systematically improved performance from roughly 40 percent correct responses in the 0 semitone condition to more than 80 percent correct responses in the 12.4 semitone condition. Indeed, performance with an alternated masker separated by 12.4 semitones from the target was actually better than with a continuous masker at some interruption rates. The result suggest that in the alternated conditions, the pitch cue was the only viable cue for segregating the target and masking speech signals, and thus it provided a substantial benefit in allowing the listener to segregate the target and masking speech signals into two streams. At these rates of interruption, listeners reported hearing two streams : a low-pitch stream, comprised of the target phrase, and a high-pitch stream comprised of the masker phrase. It is plausible that in the 12.4 semitone case at 8-16 Hz repetition rates, the pitch cue may have been so effective for segregating the target and masker into separate streams that it allowed the listeners to take advantage of the improved local SNR value in the alternating masker condition, thus explaining the higher performance obtained in that condition (though it is worth noting that performance never approached the near 100 percent correct level obtained for an interrupted speech signal with no masker).

At the highest interruption rates (64-128 Hz), there seemed to be a drop-off in performance in the alternated speech masker conditions with 5.2 and 12.4 semitone differences. The reason for this drop off is not clear, but one possible explanation is that these alternation rates were high enough to directly interfere with the perception of the F0 values of the speech signals and that they thus disrupted the segregation of the target and masking speech signals on the basis of pitch.

Except at the largest pitch separation tested, target identification in the presence of a continuous speech masker was higher than in masker interrupted conditions. A three-factor, within subject ANOVA was performed on the arcsine-transformed mean scores to verify the results described previously. The factors were masker type (continuous vs. interrupted), interruption rate, and pitch separation. Results of the ANOVA indicated a main effect of masker type ($F_{(1,7)}=53.60$, $p<0.001$), pitch ($F_{(3,21)}=55.63$, $p<0.001$), and interruption rate ($F_{(6,42)}=104.20$, $p<0.001$). All two-way interactions were significant, as was the three-way interaction of pitch, interruption rate and masker type ($F_{(18,126)}=2.39$, $p<0.001$). The significant three-way interaction was attributed to the better performance with a masker that was alternated with a target at a rate of 8 Hz and separated from it by a pitch difference of 12.4 semitones.

Although there are clearly differences between the segregation of alternating speech signals and alternating tones, it is perhaps worth mentioning that there are some parallels between the stream segregation effects seen in this experiment and those in which stream segregation with tones was examined. Experiments examining the perception of alternating high and low frequency tones have also reported a profound shift between the perception of a single alternating-frequency stream and two separate high and low frequency streams when the pitch differences exceeds a certain threshold for a given alternation rate (van Noorden, 1977; Ciocca and Bregman, 1987). Perhaps it is surprising that listeners could not derive more benefit from the pitch-segregated target and masker streams at these high rates of interruption. However, such performance decrements at high rates has been previously reported for noise maskers as well (Miller and Heise, 1950; Dirks and Bower, 1970) and has been attributed to increased non-simultaneous masking during the noise intervals. For instance, (Dirks and Bower, 1970) compared intelligibility under two conditions: speech alternating with silent intervals and speech alternating with noise. The difference in performance between the two conditions yielded a measure of the amount of non-simultaneous (forward and backward) during the noise intervals. Their results suggested that the amounts of non-simultaneous masking were more pronounced at rates of 100 Hz than 10 Hz and negligible at 1 Hz. In addition, Miller and Licklider (1950) attributed decreased intelligibility at high alternating rates to the introduction of side-bands.

2.3 Conclusions

These experiments have shown that listeners perform much differently when they are asked to extract an interrupted target speech signal from an interrupted or noise masker than when they are asked to extract the same speech signal from an interrupted or continuous speech masker. This in itself is not a surprising result. Previous research has shown that, while continuous noise is more efficient than speech in terms of its ability to energetically mask the acoustic elements of a target speech signal, a perceptually similar speech masker often produces lower performance scores because it is so much harder to segregate from the target speech. However, when an additional cue is introduced that reduces the perceptual similarity of the target and masking speech signals, performance generally increases substantially. In most experiments that have explored these similarity-based "informational masking" effects in speech-on-speech masking, similarity has been operationally defined as "sameness" in pitch (Freyman et al., 1999; Brungart, 2001; Brungart et al., 2001), vocal tract size (Darwin et al., 2003), apparent location (Freyman et al., 1999), intensity (Brungart, 2001), or some other conventional parameter of speech. In this paper, Miller and Licklider's (1950) groundbreaking work in interrupted speech perception has been extended to introduce another cue that listeners could utilize to segregate the target and masker - masker continuity.

In general, the ability to segregate a target interrupted at various rates depended largely on the type and continuity of the maskers. For noise maskers, performance was worse with a continuous than an alternating masker, presumably due to increased energetic masking. Similarly, increased performance with an alternating masker was due to a release from ener-

getic masking, so that listeners were able to "glimpse" the target during epochs when there was no masker present. On the other hand, with speech maskers, performance was best when the masker was continuous instead of alternating. Listeners were unable to *hear out* the target in the alternating masker condition, despite near infinite SNR in those conditions. We believe that performance was worse in the alternating speech masker condition because that condition eliminated some of the phonetic information in the target signal but did nothing to make the target less "similar" to the masker. The interruption of the target might also have interfered with the prosodic continuity cues that normally would have helped the listeners segregate two phrases spoken by the same talker. Conversely, in the continuous masker condition, listeners could utilize the the continuity of the masker to disambiguate it from the interrupted target speech signal, leading to a marked improvement in performance at high interruption rates.

When pitch differences were introduced between the target and masker sentences, performance was no longer improved relative to the 0 Hz baseline in the continuous masker condition, presumably because the pitch cue provided a more salient means of segregating the target and masker than the continuity cue. The pitch cues did, however, improve performance with alternating speech maskers, mainly because the pitch cue provided the listeners with a means to track the target sentence and/or ignore the masker. When rates of interruption increased, performance declined, perhaps due to interference between the interruption rates and the F0 values of the talkers.

Overall, the results from these experiments support the observation that target-masker similarity can often play a dominant role in speech-on-speech masking, and they provide additional evidence that performance with a similar-sounding speech masker cannot be predicted from the results of experiments that have used random noise as a masker. From a more practical standpoint, they also seem to rule out the possibility of using temporal alternation as a way to increase a listener's ability to process multiple speech signals through a monaural channel. If one were to look at Miller and Licklider's (1950) results with an alternating noise masker, it would be easy to become convinced that listeners who are able to understand a speech signal interleaved with noise with 100 percent accuracy might also be able to understand two interleaved speech signals with 100 percent accuracy. However, the results of this experiment clearly demonstrate that such performances are not feasible, because the listener has no way of effectively segregating the fragments belonging to the target and interfering speech signals. Yet, in real world listening environments, listeners are constantly required to reconstruct audio signals from a relatively sparse distribution of spectrotemporal "glimpses" where the target signal locally dominates the other components of a complex auditory mixture (Cooke, 2006). A more complete understanding of the cues listeners use to successfully piece together the sparsely-distributed spectrotemporal fragments of a sound into a coherent audio image is necessary before we can begin to fully understand how human listeners are able to robustly segregate audio signals in such a wide range of listening environments.

3.0 Simultaneous segregation: Assessing the contribution of call-sign to performance in multitalker listening tasks

3.1 Introduction

In tasks that require a listener to monitor multiple simultaneous voices, it is not surprising that significant declines in performance tend to occur when the number of masking talkers increases (Miller, 1947; Pollack and Pickett, 1958; Carhart et al., 1969; Bronkhorst and Plomp, 1992; Yost et al., 1996; Brungart et al., 2001; Freyman et al., 2004). While even a single competing talker can have an impact on the intelligibility of a target speech signal, the addition of a second competing talker reduces performance to a much greater degree. Previous studies that have examined multitalker listening perfomance as a function of the number of interfering talkers have found an abrupt decrease in intelligibility for the three-talker condition, and much smaller decreases in performance as additional talkers beyond the third talker are added to the stimulus. For instance, Simpson and Cooke (2005) showed that the first interfering talker reduced performance in a CVC recognition task by 15 percent, and the second interfering talker decreased performance by an additional 17 percent, but the third talker reduced performance by an additonal 8 percent. Similarly, Miller (1947) found that the addition of a second interfering talker increased the amount of masking by 8 dB relative to the single interferer case, but a third and fourth talker reduced performance by only an additional 3 dB. Finally, data from Brungart et. al., (2001) indicate that the the addition of one interfering talker reduced the number of correct target identifications by about 18 percent, and that the addition of a second interfering talker reduced performance by an another 14 percent, but that the addition of a third interfering talker decreased performance by less than 3 percent relative to the two-interferer condition [1].

All of these studies show that performance decreases dramatically when the second interfering talker is added to a multitalker stimulus, but that performance decreases much more slowly when additional interfering talkers beyond the first two are added to the ensemble. This suggests that the extraction of information from a target speech signal may somehow be a fundamentally harder task when there is more than one interfering voice present in the stimulus. Durlach (2006) has referred to this increase in difficulty as a "multimasker penalty."

Although many researchers have commented on the multimasker penalty, its origins remain unclear. Of particular interest is the question of whether the penalty is simply caused by an increase in spectral overlap between the target and masking signals (i.e., "energetic masking") or whether the penalty is more closely related to the confusability of the target speech with similar-sounding speech maskers (i.e.,"informational masking"). Certainly there is evidence that energetic masking effects caused by a reduction in "dip listening" make a

[1]The numbers for the speech-masking conditions represent mean performance in the same-talker, same-sex, and different-sex conditions of the experiment averaged across all target-to-masker ratios (TMRs) greater than or equal to -3 dB. The -3 dB lower limit was chosen in order to eliminate the impact of floor effects at low SNR values in the 3- and 4-talker conditions.

substantial contribution to the multimasker penalty. Festen and Plomp (1990) suggested that listeners can utilize momentary fluctuations in the temporal envelope of a masker to hear a target signal in the dips of the masker. As more maskers are added, these spectrotemporal gaps are filled, and thus the dip-listening strategy becomes less effective. When an infinite number of masking voices are added to the stimulus, the masker reduces to a continuous speech-shaped noise signal and no further advantages can be gained from listening in the gaps of the masking waveform. At least three studies have used noise maskers modulated with the envelope of a speech signal to examine the effects that the number of masking voices have on dip listening ability, and these studies do tend to show that the addition of the second modulated noise masker produces substantially more interference than the addition of the third and subsequent modulated noise maskers (Bronkhorst and Plomp, 1992; Simpson and Cooke, 2005; Brungart et al., 2001).

However, there is also some evidence that the multimasker penalty cannot be explained by energetic masking effects alone. Results from Brungart et. al., (2001) indicate that the addition of a second interfering speeech signal produces a much larger decrease in intelligiblity than the addition of a second modulated noise masker, even when the speech maskers are different-sex talkers that are relatively easy to segregate from the target speech. Further evidence against a purely energetic explanation for the multimasker penalty comes from a study by Carhart et. al., (1969) that compared performance in a condition with one speech masker and one noise masker to performance in a condition with two speech maskers. The results showed that both types of maskers produced roughly the same amount of interference when they were mixed together with the target speech in isolation, but that the speech reception thresholds (SRTs) were about 3.4 dB worse when the listeners tried to identify spondees in the presence of two speech maskers than when they tried to identify speech in the presence of one speech masker and one noise masker. This result, along with the results from Brungart et. al. (2001), suggests that the confusability between the target speech signal and the two masking voices may also contribute to the multimasker penalty.

The results from these studies suggest that the type of masking signal, rather than just its presence, may play a substantial role in the multimasker penalty. In particular, there may be large differences between speech signals that are contextually similar to, and thus easily confused with, the target speech, and those that are contextually much different than the target speech. There may also be differences between different types of non-speech maskers. The current study was designed to systematically investigate the relative contributions that spectral overlap (i.e., energetic factors) and target-masker confusability make to the multimasker penalty. We hypothesized that, if the multimasker penalty is driven primarily by confusability of the target and maskers, then making one of the maskers less similar to the target should lead to a significant improvement in performance. On the other hand, if the penalty is mainly due to energetic factors, then making the second masker more or less confusable with the target should have relatively little effect on performance. In order to test these possible alternatives, the study used target phrases that were drawn from the Coordinate Response Measure (CRM) (Bolia et al., 2000) corpus. The CRM phrases use a very highly structured syntax consisting of a call-sign, a color, and a number. Because of the rigid syntactic structure, target-masker confusability could be operationally defined based on the

structure of the masking phrases. For the purposes of this experiment, a CRM target phrase was presented with two types of maskers: contextually relevant (C) maskers, comprised of competing phrases from the CRM corpus, and contextually irrelevant (I) maskers, which were speech or noise maskers that contained no color or number keywords and did not follow the same syntactical structure of the CRM phrases.

3.2 Experiment 1: Effect of number and type of maskers

3.2.1 Methods

3.2.1.1 Listeners

Ten subjects, ranging in age from 19 - 55 years served as listeners. All had audiometric thresholds within the normal range ($<$ 25 dB HL at octave frequencies between 250 - 8000 Hz) and were well-practiced in similar listening tasks. Subjects were paid for their participation in the experiments.

3.2.1.2 Stimuli

A phrase drawn from the CRM corpus and containing the call-sign "Baron" was chosen as the target phrase. This target phrase was presented with either one or two other maskers. A total of ten different types of maskers were used in the experiment (see Table 1).

Table 1: **Types of maskers and corresponding designations.**

Type of Masker	**Designation**
CRM	C
Reversed CRM	C_R
English	E
Reversed English	E_R
Spanish	S
Reversed Spanish	S_R
Dutch	D
Reversed Dutch	D_R
Speech-shaped noise	N
Modulated speech-shaped noise	M

1. CRM Masker (C): The first type of masker was a standard CRM phrase that was randomly selected from all the phrases in the CRM corpus containing a different call-sign, color and number than the target phrase. The CRM masking phrase was similar

in structure and content to the CRM target phrase, so it was considered to be the most likely stimulus to produce informational masking due to the fact that the target and masker phrases were more easily confused.

2. Irrelevant Masker (I): The second type of masker was a contextually irrelevant masker that was syntactically different from, and presumably less likely to be confused with the CRM target phrase. A total of nine different irrelevant maskers were tested. Three of them consisted of normal speech: 1) Irrelevant American English speech, generated by randomly selecting a passage with the same length as the target phrase from a 1.2 min recording of a female talker reading "The Rainbow Passage" (Fairbanks, 1960); 2) Dutch speech, randomly selected from a 2.1 min recording of a female talker (Escudero and Boersma, 2002); and 3) Spanish speech, randomly selected from a 3.3 min recording of the same female talker used for the Dutch speech (who was fluent in both languages [Escudero and Boersma, 2002]). In all cases, this irrelevant speech masker was derived from a voice that was spoken by a talker who was the same sex (but never the same talker) as the target talker. In the English, Dutch, and Spanish speech cases, one additional distinct female version and two additional distinct male versions of the utterances were generated by appropriately changing the F0 and vocal tract lengths of the original female voice recordings (Darwin et al., 2003) using the software package "Praat" (Boersma and Weenink, 1996). The F0 values for the modified passages were picked so that they fell within the normal range for female and male talkers. In cases where two copies of the same masker were presented, the two maskers would consist of two different randomly selected segments of the masking phrase spoken by two different voices that were of the same sex as the target talker.

   Four additional irrelevant maskers consisted of time-reversed speech: 4) A time-reversed phrase randomly selected from the CRM corpus; 5) A time-reversed English phrase; 6) A time-reversed Dutch phrase; and 7) A time-reversed Spanish phrase. These time-reversed phrases were generated simply be reversing the waveform for the irrelevant speech masking types.

   Finally, the last two irrelevant maskers were noise maskers. The first was steady-state speech-shaped noise (SSN), which was generated by filtering a Gaussian noise through a filter whose shaped matched the long-term spectrum of 2048 phrases from the CRM corpus. The second was envelope-modulated SSN, which was created by modulating the SSN with the envelope of a randomly selected phrase from the CRM corpus (Brungart, 2001).

Table 2 shows all the different combinations of these three types of maskers that were used in the experiment. Although a very large number of individual conditions were tested, for the purposes of discussion they are collapsed into five broad categories, and a shorthand notation is used to decribe each condition based on the composition of the masking phrase. The number of letters in the designation represents the number of talkers present in the listening condition (e.g., CC vs. CCC to represent two-talker vs. three-talker conditions), whereas the letter itself represents the type of masker (e.g., CC implies that both signals

were CRM phrases and CE implies that the target was a CRM phrase and the masker was an irrelevant English phrase). The target was always a CRM phrase and always indicated by the first letter. The five categories tested consisted of all two- and three-talkers combinations containing a single CRM target phrase and one or two relevant or irrelevant maskers. Thus, the five conditions tested are refered to as CC, CI, CCC, CCI, and CII.

Table 2: **Number and type of maskers in Experiment 1.**

Notation	**Masker Designation**
CC	CC
CI	CC_R, CE, CS, CD, CM, CN
CCC	CCC
CCI	CCC_R, CCE, CCE_R, CCS, CCS_R CCD, CCD_R, CCM, CCN
CII	CC_RC_R, CEE, CEE_R, CSS, CSS_R CDD, CDD_R, CMM, CNN

All the stimuli used in the experiment were resampled to 11025 Hz, which was the lowest sampling frequency of the voice recordings used to generate the masking stimuli. Each masking stimulus was normalized to have the same overall root-mean-square (RMS) level, and the RMS level of the target signal was then varied relative to this level to set the target-to-masker ratios (TMR) to -8, -4, 0, 4 or 8 dB.

3.2.1.3 Procedure

The data were collected from listeners seated in a quiet room. All signals were presented diotically to the listeners through BeyerDynamics DT 990 Pro headphones at a comfortable listening level (approximately 65 dB SPL). On each trial, listeners heard a target sentence presented in conjunction with either one or two maskers. The listeners' task was to identify the color-number combination present in the target sentence (as identified by the phrase containing the call-sign "Baron") and to respond by selecting the appropriate colored digits displayed as a matrix of 4 colors and 8 numbers on a computer screen. Correct answer feedback was provided on every trial. For all listeners, target identification data were collected first with speech maskers and then with noise maskers. Within each block of trials, the number of maskers, the types of the maskers, and the TMR values were selected randomly from trial to trial. A total of 30 trials was collected from each subject at each TMR value in each of the twenty-seven unique masking conditions.

3.2.2 Results and Discussion

3.2.2.1 General Results

Figure 5 depicts the percent correct identification of the target color-number combinations in the two-talker conditions of the experiment. The dark gray region shows the range of performance obtained across the six different masker types comprising the CI condition. Although there was substantial variation across the different CI conditions at the -8 dB TMR value, all of the CI masker conditions performed similarly at TMR values of -4 dB and above. When the TMR value was 0 dB or better, none of the I maskers produced any significant degradation in performance in the CRM task. In the CC condition, performance was also quite good, plateauing at approximately 80 percent correct color and number identifications at negative TMR values. Thus it can be concluded from Figure 5 that the presence of a single contextually-irrelevant masker only has a significant impact on CRM performance when it is presented at a substantially higher level than the target phrase.

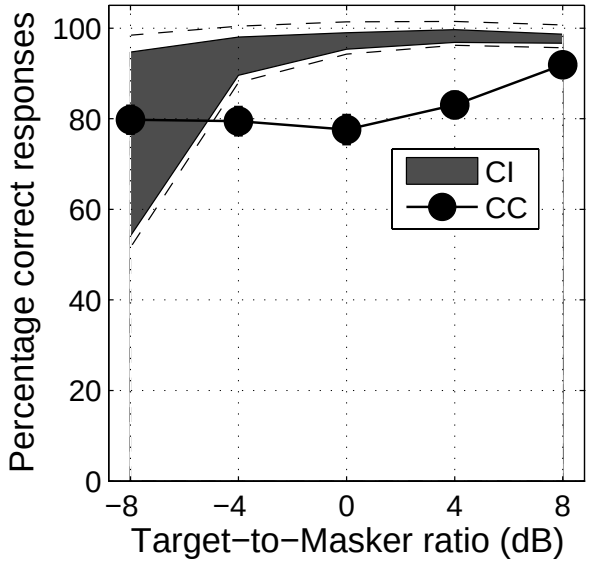

Figure 5: Percent correct identification as a function of target-to-masker ratio (TMR) for the single-masker listening condition. The circles represent target intelligibility in CC condition, with the error bars representing the 95 percent confidence estimates for the data. The gray area represents the range of performance across the five different CI conditions, with the upper bound of the region representing the best level of performance obtained in any single condition and the lower bound representing the worst level of performance obtained in any single condition. The lines bounding the gray region represents the 95 percent confidence limits for the best and worst individual conditions.

Figure 6 shows performance in the three-talker conditions of the experiment. In this figure, the darkly shaded region shows performance across the nine different masker combinations in the CII condition, the lightly shaded region shows performance across the nine

different masker combinations in the CCI condition, and the filled circles shows performance in the CCC condition. Four important features are evident from the data in the figure.

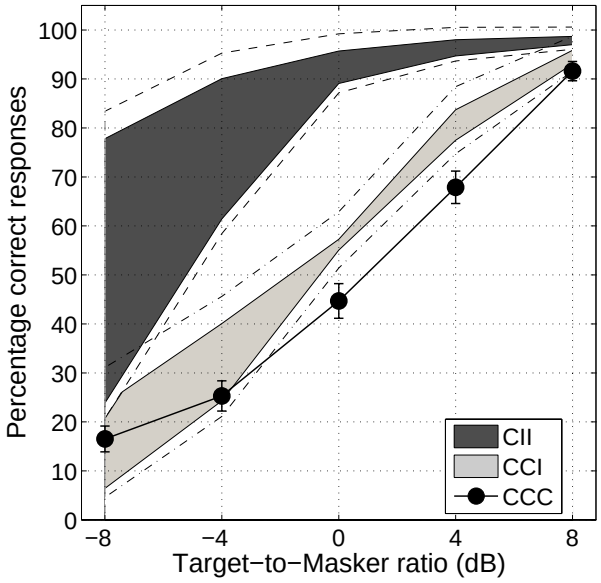

Figure 6: Percent correct target sentence identification as a function of target-to-masker ratio (TMR) for the three-talker listening conditions. The circles represent intelligibility in the CC condition, with error bars representing the 95 percent confidence intervals around the mean. The light gray area represents the range of performance in the CII conditions, whereas the dark gray region represents the range of performance in the CCI conditions. The two regions are bounded by lines that represent the 95 percent confidence intervals.

There is no evidence of a multimasker penalty in the CII condition: Although there was a wide variation in performance across the CII maskers at low TMR values, these differences narrowed as the TMR increased, and when the TMR value was 0 dB or higher performance was uniformly good across all nine CII masker combinations tested. Thus we can conclude that there is no evidence of a multimasker penalty when a target speech signal is masked by two contextually-irrelevant maskers.

There is a substantial multimasker penalty in the CCC condition: In the three-talker CCC condition, performance was substantially worse than in the two-talker CC condition. In Figure 5, performance in the CC condition never dropped below 78 percent even when the TMR was -8 dB. In contrast, performance in the CCC condition was only 45 percent when the TMR value was 0 dB and fell to 17 percent correct responses when the TMR value was -8 dB. This indicates that the addition of a second interfering CRM phrase causes a substantial decline in performance over a single interferer case, which strongly suggests the existence of a multimasker penalty in the CCC configuration.

The multimasker penalty in the CCI condition is nearly as large as in the CCC condition: Intuitively, one might expect performance in the CCC condition to be substantially worse than performance in the CCI condition because of the increased possiblity of confusing the target color and number keywords with those spoken by the second

interfering masker. However, the results in Figure 6 indicate that performance in the CCC condition was only modestly worse than performance in the CCI condition. Thus it appears that a substantial multimasker penalty occurs when at least one of the interfering talkers is contextually similar to the target speech signal.

There was little variation in performance across the different contextually-irrelevant maskers in the CCI condition. One of the most remarkable aspects of the data shown in Figure 6 is the extremely narrow range of performance that was obtained across the nine different types of 'I' maskers tested in the CCI condition. When the TMR was 0 dB, these nine maskers all resulted in performance that varied only a 3 percent range (55 percent to 58 percent). This result suggests that the specific acoustic characteristics of the 'I' masker have *little or no impact* on the amount of interference produced in the CCI listening configuration.

3.2.2.2 Masking efficiencies of the different maskers

When the TMR value was 0 dB or higher, there was relatively little variation in performance across the different 'I' maskers in the CII condition (dark region in Figure 6). However, when the TMR was -8 dB, the performance ranged from 25 percent to nearly 80 percent across the different CII configurations. This suggests that the amount of spectrotemporal overlap between the 'I' maskers and the CRM target varied substantially across the different 'I' maskers, and that these differences only emerged at low TMR values where performance was driven primarily by energetic rather than informational masking effects.

Figure 7 shows how performance at the -8 dB TMR level varied across the nine different masker combinations tested in the CII condition (see Table 1 and Table 2). These results cluster into four distinct groups. The group that resulted in the best overall performance was comprised of the six combinations of forward and reversed Dutch or Spanish speech (depicted in white bars). The next group (shown in light gray bars) included the three combinations of forward or reversed English speech. The next group (dark gray bars) consisted either of time-reversed CRM phrases or noise maskers that were modulated to match the envelope of CRM phrases. The final and by far the worst performing was the continuous noise masker (black bar). From these results, it is apparent that there were systematic differences in terms of the masking efficiencies of the different speech and noise materials used to generate the 'I' maskers. Also, not surprisingly, the continous noise masker (NN) was the most efficient masker of a speech signal. Note that the conditions containing continous noise maskers account for all of the overlap between the CII and CCI conditions seen in Figure 6. When the CII masker contained two irrelevant maskers with speech-like fluctuations in overall envelope, performance in the CII condition was consistently much better than in the corresponding CCI condition.

3.2.2.3 Discussion

The high level of performance obtained in the CII conditions of Experiment 1 demonstrate that the multimasker penalty is *not* a problem that is inherent in all speech signals where the target talker's voice is masked by more than one interfering talker. Rather, it seems to

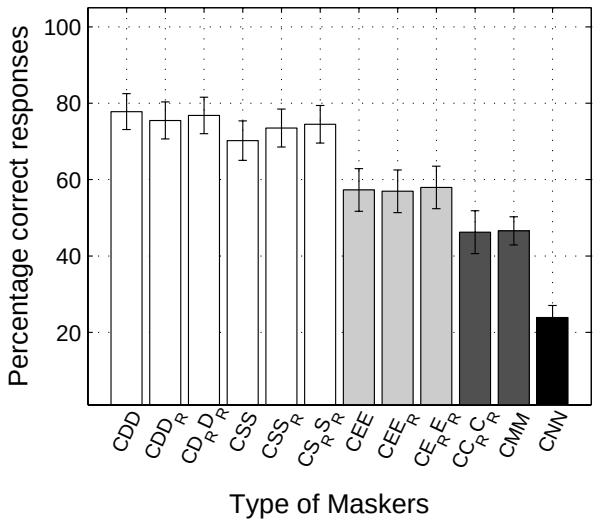

Figure 7: Percent correct target sentence identification as a function of the nine different types of maskers presented at a target-to-masker ratio (TMR) of -8 dB in the CII conditions (one target sentence presented with two irrelevant maskers). The error bars represent the 95 percent confidence intervals for the data.

be limited to the CCI and CCC conditions where at least one of the masking talkers was a CRM phrase that was contextually confusable with the target speech.

The results also indicate that the multimasker penalty is not limited to 'I' maskers that are perceptually similar to masking speech. Indeed, when the TMR value was 0 dB or above, performance in the CCI condition was little different for 'I' maskers that consisted of speech, reversed speech, modulated noise, or continous noise. This seems to indicate that the multimasker penalty specifically occurs in situations where the listener is required to segregate two contextually similar speech signals in the presence of *any* kind of spectrotemporally overlapping energetic noise masker.

A more complete understanding of the nature of this multimasker penalty can be obtained by more carefully analyzing the types of errors that occured in the experiment. Under ideal conditions, the task of the listener in the CRM task is to listen for the target talker who says the "Baron" call sign, and follow that talker's voice through the phrase to determine the target color and number keywords. However, in cases where this goal cannot be achieved, listeners are likely to revert to a backup strategy of trying to extract one or more valid color and number keywords in the stimulus and determine which one was spoken by the target talker. Although the listeners were not always able to correctly identify the target keywords, they were consistently able to identify at least one color keyword and at least one number keyword in every stimulus condition tested. In the CII condition, where there were no competing color or number keywords, listeners were able to correctly identify the target keywords more than 90 percent of the time when the TMR was 0 dB. Since there was no need to connect the color and number keywords to the call sign in the CII condition, the ability to successfully extract a valid color and a number from a three-talker stimulus was

sufficient to achieve a high level of performance in that condition. In the CCC condition, where there were three sets of colors and numbers, fewer than 2 percent of the responses contained a color or number keyword that was not spoken by one of the CRM talkers in the stimulus. However, in that condition, correct identification of the target keyword required the listeners to determine which color-number keyword was spoken by the talker who used the "Baron" call sign. The listeners did not seem to be as good at that part of the CRM task, either because they really only heard one set of color and number keywords, or because they heard all three sets of keywords and were unable to successfully link the call-sign and color-number portions of the target phrase together into a single response.

In order to differentiate between these two possibilities, it is really necessary to find some way to determine whether the listeners were in fact able to hear more than one set of color and number keywords in the stimulus. One way to investigate this possiblity is to generate a masking phrases that are syntactically identical to the CRM phrases, but contain "invalid" color and number key words that are not contained in the CRM response set. If the multimasker penalty found in Experiment 1 occurred because the listeners were able to hear all the color-number keywords but were unable to determine which one was spoken by the target talker, then performance in this experiment would be expected to be much higher than in Experiment 1 (because the listeners could identify the target keywords as the only valid color and number words contained in the stimulus). However, if the limiting factor in Experiment 1 was the ability to hear all of the color-number keywords simultaneously, then performance should be comparable to Experiment 1, with a somewhat larger number of random response errors not matching the color-number keywords present in the stimulus. Experiment 2 was conducted to explore these possiblities in more detail.

3.3 Experiment 2: Syntactically similar CRM maskers

One of the remarkable outcomes of Experiment 1 was that there did not appear to be any difference across the different irrelevant maskers (I) in terms of the amount of interference they caused in the various CCI condition. However, there must be some limits on how similar to the target an 'I' masker can be before it starts to have a larger negative impact on performance in the CCI conditions. In Experiment 2, the 'I' masker was replaced with a masking phrase that had a syntactic structure that was identical to the CRM phrases, keywords that were similar to the target keywords, but were not part of the target response set. These out-of-set masking phrases, represented by 'O' in the stimulus condition set, contained two colors and two numbers; the two colors were "black" and "brown" and the two numbers were "nine" and "ten."

3.3.1 Methods

Ten listeners participated in the experiment, seven of whom also participated in Experiment 1. Recordings of the out-of-set corpus were obtained from four male and four female talkers. For conformity, recordings of all the phrases in the original CRM corpus were also obtained from these talkers. The out-of-set maskers were denoted by the letter 'O', while the normal "in-set" CRM phrases were designated by the letter 'C' (as in Experiment 1).

As a control condition, a contextually irrelevant masker consisting of a time-reversed phrase randomly selected from the out-of-set corpus was also tested and designated by the letter 'I'. Thus, the listening conditions of Experiment 2 were : CC, CCC, CCO, CCI, COO, CII, and COI. All stimuli were presented diotically to listeners at a TMR of 0 dB. As in Experiment 1, the response matric depicted an array of four colors and eight numbers, and did not contain the 'O' colors and numbers.

3.3.2 Results and Discussion

3.3.2.1 General Results

The performance in each listening condition is depicted in Figure 8. The dark gray bar towards the top of the figure depicts mean performance ± 1 standard error in the CC condition, whereas the light gray bar towards the bottom depicts mean performance ± 1 standard error in the CCC condition. It is worth noting performance in the control conditions was somewhat lower in Experiment 2 than Experiment 1; the differences could be attributed to greater similarity in the talkers in the new corpus recorded to Experiment 2, or it might be attributed to the fact that the listeners in the current experiment had extensive practice with the talkers in the original CRM corpus, but were only minimal experienced with the talkers in the new recordings.

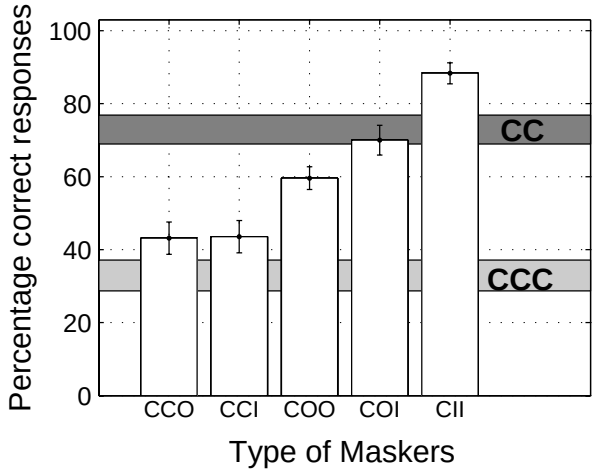

Figure 8: Percent correct target sentence identification as a function of the type of maskers. The two horizontal bars represent mean performance ± 1 standard error with one CRM masker(dark gray bar towards the top) and two CRM maskers(light gray bar towards the bottom) respectively. The bars represent performance in the following listening conditions : CCO, CCI, COO, COI, and CII.

The major findings of Experiment 2 can be summarized as follows:

Performance in the COO condition was substantially better than performance in the CCC condition. In the COO condition, listeners were able to correctly identify

both the color and number in the target CRM phrase in 60 percent of the trials. If one assumes that the "black" and "brown" color keywords and the "nine" and "ten" number keywords in the 'O' maskers produce roughly the same amount of masking as the other color and number keywords in the CRM corpus, then this result implies that listeners have a 60 percent chance of hearing both the color and number keywords in any one of the three CRM phrases contained in the CCC stimulus. If one accounts for the additional processing required to discriminate between the valid and invalid keywords in the COO condition, then it is likely that even this 60 percent figure underestimates how intelligible the individual voices are in the CCC stimulus. Thus it seems that the relatively low levels of performance achieved in the CCI and CCC conditions of Experiment 1 cannot be explained simply by the inaudibilty of the target CRM phrase in the three-talker stimulus.

Performance was nearly identical in the CCI and CCO conditions. Both of these conditions produced slightly higher performance than the CCC condition, just as the CCI condition produced slightly higher performance than the CCC condition of the first experiment. But there was little or no difference in overall performance across the CCI and CCO conditions. This somewhat surprising result further emphasizes the finding from Experiment 1 that the specific characteristics of the irrelevant masker have virtually no impact on the amount of interference caused in the CCI condition. In Experiment 2, the 'O' masker had exactly the same syntax as the CRM target, differing only in terms of the specific color and number keywords it contained. In contrast, the 'I' masker was a time-reversed speech waveform that had almost nothing in common with the CRM target. Yet both maskers produced almost exactly the same degradation in overall performance when they were added to the CC condition.

Another important implication of the equivalent performance seen in the CCI and CCO conditions is that call sign audibility probably also had a relatively small impact on performance in the CCC condition. If performance in the CCC condition were limited by the listener's ability to hear the target call-sign, then one would expect somewhat lower performance in the CCO condition, where all three talkers spoke valid call signs from the CRM corpus, than in the CCI condition, where only two call signs were present in the stimulus. Thus, although call sign audibility was not explicitly tested, it is likely that listeners were at least as good at hearing out the target call sign in the three talker CRM ensemble as they were at hearing out the color and number keywords.

3.3.2.2 Error Analysis

The relatively high level of performance achieved in the COO condition of Experiment 2 strongly suggests that listeners are able to hear at least two and possibly all three of the voices in a CCC stimulus the majority of the time. A more detailed analysis of this ability is provided by the error analysis shown in Figure 9. This figure shows the percentage of "random" color and number responses that did not match any of the CRM phrases in the stimulus for each individual condition in Experiments 1 and 2. If we assume that listeners only made random responses when they failed to hear a valid color and number keyword in the stimulus, then we can see from the COO results that the listeners were able to hear a single valid color keyword in a three-talker stimulus roughly 70 percent of the time, and

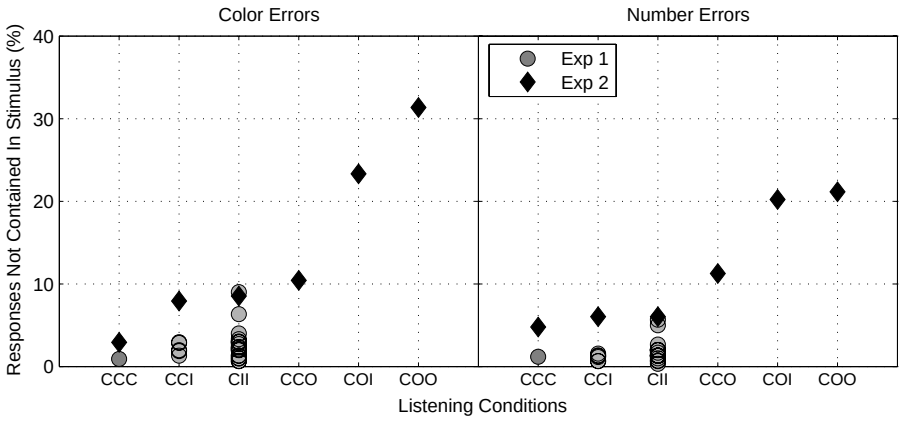

Figure 9: Percent number (left panel) or color (right panel) errors as a function of listening conditions in Experiment 1 and 2. Data points from Experiment 1 are depicted by circles and those in Experiment 2 by diamonds.

a single valid number keyword in a three-talker stimulus roughly 80 percent of the time. From the CCO results, we see that the listeners were able to hear at least one out of two valid keywords roughly 90 percent of the time. And from the CCC results, we see that they were able to hear at least one out of three valid keywords 95-98 percent of the time. At the end of last section, we postulated that poor performance in the three-talker condition might occur for either of two reasons. The first possiblility was that the listeners were only able to hear the keywords spoken by one of the three talkers in the stimulus. This possiblity has been largely discounted by the results of Experiment 2. The second possiblity was that the listeners were generally able to hear the call signs, colors, and numbers spoken by the competing talkers in the CCC stimulus, but that they were unable to reliably attribute the target color and number keywords to the phrase containing the "Baron" call sign. In light of the results of Experiment 2, this appears to be the more likely possiblity. However, in cases where comparisons are made between CRM and non-CRM maskers, the overall percentage of correct color and number identifications may not be the ideal metric for assessing the ability to tie the target keywords together with the target call sign. The overall correct identification metric does not account for cases where a listener was unable to use the target call sign to follow the target phrase but still obtained a correct response by randomly guessing among the valid color and number keyword combinations heard in the stimulus. Such a tactic would have a higher probability of success in the CCI condition than the CCC condition, because of the smaller number of valid keywords in the stimulus. The next section introduces a new metric of performance in the CRM task designed to account for the effects of this type of guessing and give a more accurate estimate of how well the listeners were able to use information contained in the target call sign to perform the CRM task.

3.4 Assessing the contribution of call-sign to performance in the CRM task

The results of Experiment 2 showed that listeners are generally able to hear more than one valid color and number keyword combination in a CCC or CCI stimulus. Thus there is a possiblity that they might sometimes get a correct response not by determining which color and number keyword was spoken by the target talker, but rather by randomly guessing one of the valid color-number keyword combinations heard in the stimulus. If one assumes that the listener is equally likely to hear any of the color and number keyword combinations contained in the stimulus, then it is possible to construct a metric of CRM performance that accounts for the impact of these random keyword guesses and gives a more direct assessment of the impact that the call-sign portion of the stimulus has on overall performance in the CRM task.

At the heart of the calculation is the assumption that the color and number keywords spoken by all the CRM talkers in the stimulus are equally likely to be heard by the listener, and that a response that matches a talker other than the target talker is in effect a random guess that represents a trial where the listener was not able to successfully make use of the information contained in the target call sign. If one assumes that listener who is uncertain about the identity of the target talker is equally likely to respond with any color-number combination contained in the stimulus, then it follows that there should be just as many trials where this listener accidentally selected the correct target keywords as there are trials where the listener selected any of the other masking voices in the stimulus. Thus, the percentage of correct responses can be "corrected" for guessing by subtracting the probability of randomly selecting one of the maskers from the overall probability of correct responses in the 0 dB TMR condition. In other words,

$$\gamma = Correct\ responses - \frac{Responses\ matching\ masker}{Number\ of\ maskers}$$

where γ is the percentage of correct responses *corrected for guessing*. (Note that this calculation is only valid when the TMR is 0 dB, because it relies on the assumption that the listener has no reason to favor one talker over another in the selection of a random response.) Roughly speaking, the corrected response γ can be interpreted as the true percentage of time when the listeners were effectively able to make use of the call-sign "Baron" to extract the target speech from the stimulus. In the limit, a γ value of zero would indicate that the listeners were gaining no advantage from the presence of the call-sign. This would be the case if, for example, the listeners were only able to hear one dominant color and number keyword pair in the three-talker stimulus and responded with that keyword pair.

Table 3 shows the results of the correction calculation for the number keyword responses of Experiments 1 and 2. In Experiment 1, γ was calculated to be 70 percent in the condition with one CRM masker, suggesting that the listeners were able to make effective use of the call-sign in the 2-talker condition. In the three-talker CCC condition, γ dropped to 39 percent, indicating that, in contrast to the relatively modest effect a third talker has on extraction of the CRM keywords, the addition of a third talker has a dramatic effect on

the ability of a listener to effectively tie the keywords back to the call-sign in the target utterance.

Table 3: **Calculation of corrected percent correct for number keyword responses. All calculations represent data collected with a TMR value of 0 dB.**

Condition	Number of Maskers	Correct Responses	Responses Matching	Corrected Percent Correct (γ)
Experiment 1				
CC	1	0.85	0.14	**0.7**
CCC	2	0.59	0.39	**0.39**
CCI	2	0.68-0.70	0.28-0.30	**0.39-0.42**
Experiment 2				
CC	1	0.80	0.18	**0.62**
CCC	2	0.44	0.56	**0.18**
CCI	2	0.56	0.38	**0.18**

The most striking aspect of the table is the fact that, across all the nine different combinations of one CRM masker and one irrelevant masker (CCI), the calculated value of γ for number keywords only varied from 0.39 to 0.42. Furthermore, this figure almost exactly matches the value of γ for the CCC condition with three CRM talkers. This suggests that there was essentially no difference in the ability of the listeners to effectively use the target call-sign to identify the target keywords in the CCC and CCI conditions. Rather, it seems the difference in performance between the two conditions can be explained almost entirely by the fact that listeners who were unable to associate a color-number pair with the target call sign, and made a random guess from among the keywords in the stimulus had a 50 percent chance of accidently identifiying the target talker in the CCI condition, but only a 33 percent chance in the CCC condition.

The bottom of Table 3 shows the same calculations from the conditions of Experiment 2. In both the two-talker and three-talker conditions, overall performance in Experiment 2 was lower than in Experiment 1, presumably because the listeners were less familiar with the talkers in the corpus. In fact, when corrected for guessing, performance γ in the three-talker conditions of Experiment 2 was reduced more than 50 percent relative to Experiment 1, from 39 percent to 18 percent. However, as in Experiment 1, it is the case that the corrected score γ is almost identical for the case with two CRM maskers (CCC) and the case with one CRM masker and one irrelevant masker (CCI). The same pattern is also repeated in Table 4, which shows the calculation of γ for the color keywords.

3.5 Discussion and Conclusions

The results of the γ metric, together with the results from Experiments 1 and 2, demonstrate that the multimasker penalty incurred by the adddition of a second masker (relevant or

Table 4: **Calculation of corrected percent correct for color keyword responses. All calculations represent data collected with a TMR value of 0 dB.**

Condition	Number of Maskers	Correct Responses	Responses Matching	Corrected Percent Correct (γ)
Experiment 1				
CC	1	0.82	0.16	**0.66**
CCC	2	0.57	0.42	**0.36**
CCI	2	0.66-0.68	0.28-0.32	**0.35-0.41**
Experiment 2				
CC	1	0.79	0.18	**0.61**
CCC	2	0.45	0.51	**0.17**
CCI	2	0.55	0.38	**0.17**

irrelevant) in the CRM task is the result of a reduction in the listener's ability to tie together the call-sign, color, and number keywords in the target CRM phrase. There is evidence that listeners are still able to perform *simultaneous* segregation tasks, such as the extraction of a color word or a number word from a multitalker mixture, in the CCI condition. But their ability to perform *sequential* grouping tasks, such as tying together words spoken by a single talker at different times in a sentence, appears to be severely impaired.

In cases like the CII condition tested here, where both maskers are semantically and syntactically very different from the target phrase, there is no evidence of a multimasker penalty. This is probably true because the listener has enough context to "know what to listen for" at any given point in the stimulus and thus can accomplish the task by listening for the target phrase continuously throughout the mixture. However, much more difficulty occurs in cases where there is some ambiguity about the content of the target phrase. In these cases, listeners have to rely on their ability to use prosody, talker characteristics, and other acoustic features to follow the target phrase in a continuous signal that contains multiple talkers. It is this ability that appears to degrade markedly in cases where the target sentence is masked by more than one competing sound.

Remarkably, the amount of interference caused by the second interfering voice in a three-talker stimulus is almost completely independent of the acoustic characteristics of that voice, at least in cases where the TMR is near 0 dB. In this experiment, very similar results were found in the CCI condition when the 'I' masker was as different from the target phrase as continous noise and as similar to the target phrase as a CRM phrase with out of set color and number keywords.

At this point, the exact reason for the multimasker penalty is not yet completely understood. It might simply be related to a limitation in attentional resources. When the task requires the listener to separate more than two simultaneous contextually similar voices, the basic workload of extracting phonetic information from the stimulus increases. This may

draw away working memory or other attentional resources that might otherwise be deployed to help focus on a target talker's voice over time and link the elements of the target phrase into a coherent stream of words.

Another possible effect that the second masker might have on performance is the elimination of redundant acoustic information in the target signal that is unnecessary for extracting the target speech from a non-confusable masker but essential for extracting the targe speech from a confusable masker. In the CCI task, the listener must listen for the "Baron" call-sign, determine which talker is speaking that call-sign, and then follow the intonation and style of that talker through to the color and number portion of the CRM stimulus. It is quite possible that the addition of the 'I' masker could energetically mask out some of the acoustic cues in the stimulus that make it possible to follow the speech of the target talker. In contrast, in the CII task, it is only necessary for the listener to identify the color and number words spoken in the stimulus, without regard to who the target talker might be. The acoustic cues masked out by the second interfering talker might be largely irrelevant in this much simpler task.

The fact that the multimasker penalty has more impact on the listener's ability to tie together connected discourse than on the ability to extract target keywords may also help explain why the results of this experiment differ from those of some earlier experiments examining the multitmasker penalty. For example, some of the data from Carhart et. al. (1969) cannot be reconciled with findings from the current study. Carhart et. al. (1969) reported decreases in target recognition of approximately 2.2 dB with a single speech masker or a single noise masker, and 3.2 dB with a combination of a speech and a noise masker, but 6.6 dB with two speech maskers. Thus it is clear that, in their study, the nature of the third masker has a large effect on target recognition, whereas in our study, the nature of the third talker did not matter much. The likely explanation is that in their study, the spondee targets were not confusable with the maskers and consequently energetic masking effects dominated the results. The results of Experiment 2 suggest that the listeners were able to understand multiple simultaneous keywords in this study, and thus they were operating in a range where confusability, rather than detectability, was limiting performance.

On a final note, data from multiple studies documenting the multimasker penalty (Miller, 1947; Pollack and Pickett, 1958), the results from the current study as well as the resource limitation explanation proposed above, lend support to the "Listener-Min" listening strategy as proposed by Durlach (2006). In the two-talker condition, listeners can maximize their performance by nulling out one masker. The addition of a second masking voice reduces the effectiveness of that strategy. For the "Listener-Min" strategy or the resource limitation explanation to be true, adding a third interfering talker to a three talker mixture should result in a relatively small decrease in performance, since the masker-nulling strategy has already failed. One example of a strategy based on nulling a masker is the F0 cancellation strategy proposed by (de Cheveigne, 1993). However, it should be noted that a strategy based on F0 cancellation is likely to be much less effective in a situation where there are two interfering voices with different F0s than when there is one voiced masker and one unvoiced (noise) masker. Thus, on the surface our results do not seem to provide much support for the cancellation theory. However, it is possible that the F0 cancellation strategy also relies

on redundant acoustic cues that might be masked out by the addition of a noise masker, so it is still possible that some portion of the multimasker penalty might be related to F0 cancellation.

References

Arbogast, T., Mason, C., and Kidd, G. (2002). The effect of spatial separation on informational and energetic masking of speech. *Journal of the Acoustical Society of America*, 112:2086–2098.

Boersma, P. and Weenink, D. (1996). Praat, a system for doing phonetics by computer, version 3.4. *Institute of Phonetic Sciences, University of Amsterdam*, 134:1–182.

Bolia, R., Nelson, W., Ericson, M., and Simpson, B. (2000). A speech corpus for multitalker communications research. *Journal of the Acoustical Society of America*, 107:1065–1066.

Bronkhorst, A. and Plomp, R. (1992). Effects of multiple speechlike maskers on binaural speech recognition in normal and impaired listening. *Journal of the Acoustical Society of America*, 92:3132–3139.

Brungart, D. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of the Acoustical Society of America*, 109:1101–1109.

Brungart, D., Simpson, B., Ericson, M., and Scott, K. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *Journal of the Acoustical Society of America*, 110:2527–2538.

Buss, E., Hall, J., and Grose, J. (2004). Spectral integration of synchronous and asynchronous cues to consonant identification. *Journal of the Acoustical Society of America*, 115:2278–2285.

Carhart, R., Tillman, T., and Greetis, E. (1969). Perceptual masking in multiple sound backgrounds. *Journal of the Acoustical Society of America*, 45:694–703.

Ciocca, V. and Bregman, A. (1987). Perceived continuity of gliding and steady-state tones through interrupting noise. *Perception and Psychophysics*, 42:476–484.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America*, 119:1562–1573.

Darwin, C., Brungart, D., and Simpson, B. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *Journal of the Acoustical Society of America*, 114:2913–2922.

de Cheveigne, A. (1993). Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *Journal of the Acoustical Society of America*, 93:3271–3290.

Dirks, D. and Bower, D. (1969). Masking effects of speech competing messages. *Journal of Speech and Hearing Research*, 12:229–245.

Dirks, D. and Bower, D. (1970). Effect of forward and backward masking on speech intelligibility. *Journal of the Acoustical Society of America*, 47:1003–1008.

Drake, C. and McAdams, S. (1999). The auditory continuity phenomenon: Role of temporal sequence structure. *Journal of the Acoustical Society of America*, 106:3529–3538.

Durlach, N. (2006). Auditory masking: Need for improved conceptual structure. *Journal of the Acoustical Society of America*, 120:1787–1790.

Escudero, P. and Boersma, P. (2002). The subset problem in l2 perceptual development: Multiple-category assimilation by dutch learners of spanish. *in Proceedings of the 26th annual Boston University Conference on Language Developments*. Edited by B. Skarabela, and S. Fish, and A. H.-J. Do (Cascadilla Press/Somerville, MA).

Fairbanks, G. (1960). *Voice and articulation drillbook, Second Edition*. Harper and Row, New York.

Festen, J. and Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing. *Journal of the Acoustical Society of America*, 88:1725–1736.

Freyman, R., Balakrishnan, U., and Helfer, K. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *Journal of the Acoustical Society of America*, 115:2246–2256.

Freyman, R., Helfer, K., McCall, D., and Clifton, R. (1999). The role of perceived spatial separation in the unmasking of speech. *Journal of the Acoustical Society of America*, 106:3578–3588.

Hawley, M., Litovsky, R., and Culling, J. (2000). The 'cocktail party' effect with four kinds of maskers: Speech, time-reversed speech, speech-shaped noise, or modulated speech-shaped noise. *Proceedings of the Midwinter Meeting of the Association for Research in Otolaryngology*, page 31.

Howard-Jones, P. and Rosen, S. (1993). Uncomodulated glimpsing in "checkerboard" noise. *Journal of the Acoustical Society of America*, 93:2915–2922.

Kidd, G. J., Mason, C., Deliwala, P., Woods, W., and Colburn, H. (1994). Reducing informational masking by sound segregation. *Journal of the Acoustical Society of America*, 95:3475–3480.

Kidd, G. J., Mason, C., Rohtla, T., and Deliwala, P. (1998). Release from informational masking due to the spatial separation of sources in the identification of nonspeech auditory patterns. *Journal of the Acoustical Society of America*, 104:422–431.

Kryter, K. C. (1969). American national standards methods for calculation of the articulation index. American National Standards Institute, New York. ANSI S3.5-1969.

Mackersie, C., Prida, T., and Stiles, D. (2001). The role of sequential stream segregation and frequency selectivity in the perception of simultaneous sentences by listeners with sensorineural hearing loss. *Journal of Speech, Language and Hearing Research*, 44(1):19–28.

Miller, G. (1947). Sensitivity to changes in the intensity of white gaussian noise and its relation to masking and loudness. *Journal of the Acoustical Society of America*, 191:609–619.

Miller, G. A. and Heise, G. A. (1950). The trill threshold. *The Journal of the Acoustical Society of America*, 22(5):637–638.

Moulines, E. and Charpentier, F. (1990). Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467.

Neff, D. (1995). Signal properties that reduce masking by simultaneous random-frequency maskers. *Journal of the Acoustical Society of America*, 96:1909–1921.

Pollack, I. (1955). Masking by a periodically interrupted noise. *Journal of the Acoustical Society of America*, 27:353–355.

Pollack, I. and Pickett, J. (1958). Stereophonic listening and speech intelligibility. *Journal of the Acoustical Society of America*, 30:131–133.

Rose, M. and Moore, B. (1997). Perceptual grouping of tone sequences by normally hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 102(3):1768–1778.

Simpson, S. and Cooke, M. (2005). Consonant identification in n-talker babble is a non-monotonic function of n. *Journal of the Acoustical Society of America*, 116:2775–2778.

Thurlow, W. and Elfner, L. (1959). Continuity effects with alternately sounding tones. *Journal of the Acoustical Society of America*, 31:1337–1339.

van Noorden, L. (1977). Minimum differences of level and frequency for perceptual fission of tone sequences abab. *Journal of the Acoustical Society of America*, 61:1041–1045.

Warren, R. (1970). Restoration of missing speech sounds. *Science*, 167:392–393.

Warren, R., Obusek, C., and Ackroff, J. (1972). Auditory induction: Perceptual synthesis of absent sounds. *Science*, 176:1149–1151.

Wilson, R. and Carhart, R. (1969). Influence of pulsed masking on the threshold for spondees. *Journal of the Acoustical Society of America*, 46:998–1010.

Yost, W., Dye, R., and Sheft, S. (1996). A simulated 'cocktail party' with up to three sources. *Perception and Psychophysics*, 58:1026–1036.